Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Computers and Electronics in Agriculture 46 (2005) 153–179

Computers and electronics in agriculture

www.elsevier.com/locate/compag

Sensor-directed response surface sampling designs for characterizing spatial variation in soil properties

S.M. Lesch*

USDA-ARS, George E. Brown Jr., Salinity Laboratory, 450 W. Big Springs Road, Riverside, CA 92507, USA

Abstract

In many applied precision farming applications, remotely sensed survey data are collected specifically because these data correlate well with some soil property of interest. Additionally, a general model for the functional relationship between the soil property and the sensor data is often known a priori, but the exact parameter estimates associated with the model must still be determined via some type of site-specific sampling strategy. The main objective of this paper is to present an objective sampling and simplified modeling strategy for predicting soil property information from such spatially referenced sensor data. Some common types of spatial linear prediction models and linear geostatistical models are reviewed, and the assumptions needed to reduce these more complicated models to a spatially referenced, ordinary linear regression model (LR) are discussed. Next, a model-based sampling strategy for estimating an ordinary linear regression model in the spatial setting is described. This sampling strategy incorporates a traditional response surface design into an iterative, space-filling type algorithm for purposes of selecting sample site locations that are (i) nearly optimal with respect to matching the selected response surface design levels and (ii) physically separated far apart as possible to ensure the best chance that the independent error assumption is adequately met. This strategy can in principal be used to select a minimal number of optimal sample site locations that satisfy the residual independence assumptions in the ordinary model. A detailed case study of a salinity survey using electromagnetic induction (EM) and four-electrode sensor data is then presented. These case study results confirm that the sampling strategy was highly effective at ensuring efficient regression model parameter estimates and a reliable salinity prediction map. An additional simulation study confirmed the effectiveness of this model-based strategy over a more traditional simple random sampling strategy with respect to four regression model design criteria. Under the right conditions, this methodology should be applicable to many types of precision farming survey applications where

* Corresponding author.
E-mail address: slesch@ussl.ars.usda.gov.

0168-1699/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.compag.2004.11.004

soil property/sensor data prediction models need to be fitted using only a limited number of soil samples.

Published by Elsevier B.V.

Keywords: Precision agriculture; Salinity; Spatial linear model; Variogram; Response surface design

1. Introduction

The collection of apparent soil electrical conductivity (EC_a) survey data for the purpose of characterizing various spatially referenced soil properties has received considerable attention in the soils literature in the last 20 years (Corwin and Lesch, 2003). Most of the original interest was directed towards the characterization of field scale soil salinity patterns (Rhoades et al., 1999; Hendrickx et al., 2002). However, EC_a survey data are being increasingly used in precision farming applications in an effort to obtain information on numerous soil properties. In practice, apparent soil conductivity survey data often correlate reasonably well with various soil properties (salinity, soil texture, soil water content, etc.) under different field conditions (Lesch and Corwin, 2003). Not surprisingly, EC_a data have therefore been used extensively in precision agriculture survey applications for characterizing the spatial variability of these properties.

The basic idea behind the theory of precision farming is to exploit the knowledge of the inherent spatial variability of the soil and crop condition(s) in a specific field or region to design better management practices (Larson and Robert, 1991). In turn, these better agricultural management practices should lead to higher crop yield and/or more optimal use of agrichemicals, water, time and labor, etc., and thereby improve sustainability through increased production and profit margin, decreased input requirements, and/or a reduction in detrimental environmental impacts.

In principal, many precision farming management strategies hold great promise. However, in practice, these same management strategies are greatly affected by both the availability and accuracy of the spatially referenced input soil properties. In situations where direct, exhaustive sampling must be employed to gather adequate spatial precision about one or more input soil properties, the theoretical gain in profit from a site-specific management strategy can be quickly offset by the extra cost incurred by the sampling effort. Hence, many promising management strategies are often not cost effective in practice, due to the need for exhaustive sampling of the necessary baseline input variable(s).

This sampling cost issue has led many researchers to aggressively pursue the idea of collecting secondary ground- or air-based remote sensing information as surrogate data, i.e., data that can be used to help infer the detailed spatial pattern(s) of the primary input property(ies) of interest. Survey EC_a data are perhaps the most common example of surrogate remote sensing data, but certainly not the only one. Other examples of remote sensing data are numerous, and include various types of imagery data, natural gamma ray measurements, time-domain electromagnetic induction (EM) and time-domain reflectometry, ground-penetrating radar, and direct-yield monitoring measurements, etc. Regardless of the type of data collected, the basic idea is the same: the sensor data are acquired to increase knowledge about the underlying target soil property(ies) of interest, and thereby greatly reduce the need for acquiring baseline soil samples.

In most applied surveys, the collected sensor data are known to correlate well with one or more target soil properties of interest. Additionally, the general relationship (i.e., the model structure) can be reasonably well specified a priori, but accurate values for the model parameters can only be obtained through directed baseline sampling efforts. Thus, an obvious question arises: how should the samples be chosen (i.e., where and how many)?

The objective of this paper is to present a coherent, objective model-based sampling strategy for addressing the statistical issues of (i) when and how a spatially referenced regression model can be used to predict a target soil property from acquired survey data and (ii) where baseline calibration soil sample data should be acquired in order to optimize the model estimation process. The techniques presented here originated out of the need to accurately predict spatial soil salinity patterns from acquired EC_a survey data without relying on excessive soil sampling (Lesch et al., 1995a,b, 2000). However, the underlying statistical methodology is quite general and directly applicable to the broader precision farming sampling and modeling issue mentioned above.

2. Spatial linear prediction models

2.1. Model specification

In a typical field survey where some type of ancillary sensor data are collected, the general goal is to use the ancillary sensor data to help predict a specific, unobserved soil property. Without loss of generality, define the relationship between the soil property measurements, y, and sensor data, $\mathbf{z}$, to be

$$y = g(\mathbf{z}) + e \tag{1}$$

where $g(\mathbf{z})$ represents some unknown function of the vector of k-collocated sensor readings, and e represents a random error component. Now assume that Eq. (1) can be adequately approximated using a suitably defined spatial linear prediction model:

$$\mathbf{y} = \mathbf{Z}\beta + \eta \tag{2}$$

where $\mathbf{y}$ represents an $(m\,x\,1)$ vector of observed soil property data, $\mathbf{Z}$ represents an $(m\,x\,p)$ data matrix that includes observed functions of sensor readings collected at the same m collocated survey sites, $\boldsymbol{\beta}$ represents a $(p\,x\,1)$ vector of unknown parameter estimates and η represents a second-order stationary, jointly normal random error process. Typical stationary spatial structures for η are well documented in the geostatistical-statistical literature (Cressie, 1991; Haining, 1990; Wackernagel, 1998; Webster and Oliver, 2001); examples in two dimensions include the isotropic and anisotropic exponential, spherical, and gaussian covariance structures, either with or without nugget effects. (For the remainder of this paper the residual errors in Eq. (2) will always be assumed to follow a joint normal distribution, regardless of any additional assumptions made concerning the covariance structure.)

Eq. (2) represents a versatile linear prediction model that can incorporate various types of modeling assumptions. When the error process is assumed to be uncorrelated at all positive lag distances (η represents a pure nugget-effect), then Eq. (2) reduces to an ordinary linear regression model (LR). This model achieves all of its predictive capabilities via regression on the sensor data (**Z** matrix), since the residuals are spatially independent (under the assumption of joint normality). When a second-order stationary random error process is assumed, then Eq. (2) is commonly referred to as a spatial linear regression (SLR) model. This model incorporates both a regression component and spatially correlated residual error structure to improve the prediction accuracy. An even more comprehensive approach is the hierarchical spatial linear regression (HSLR) model, which was recently introduced by Royle and Berliner (1999). In this model, the sensor readings themselves are also assumed to exhibit some type of stochastic spatial structure, and this structure is in turn used to interpolate the sensor data (and thus the regression model predictions) across the entire spatial domain.

Many popular geostatistical modeling techniques tend to be very similar to the classical spatial prediction models discussed above. As discussed by Royle and Berliner (1999), the HSLR model can be viewed as an alternative form of a cokriging (CoK) model, developed from a hierarchical (i.e., conditional) viewpoint. Likewise, the SLR model is mathematically equivalent to a universal kriging (UK) equation and/or a kriging with external drift (KED) equation (Cressie, 1991; Rivoirard, 2002; Wackernagel, 1998). KED models are commonly used to incorporate known (observed) auxiliary covariate data into the kriging system, while UK models are commonly used when the expected value of the random function is assumed to a polynomial function of the position coordinates. In both cases the auxiliary covariate data are treated as fixed (i.e., non-stochastic) and observed.

As discussed in Cressie (1991) and originally developed by Goldberger (1962), the best linear unbiased predictor for the response variable (i.e., soil property) in the SLR, KED, and/or UK model can be expressed as

$$y_0 = \mathbf{c}'\Sigma^{-1}\mathbf{y} + (\mathbf{z}_0' - \mathbf{c}'\Sigma^{-1}\mathbf{Z})\mathbf{b} = \mathbf{c}'\Sigma^{-1}(\mathbf{y} - \mathbf{Z}\mathbf{b}) + \mathbf{z}_0'\mathbf{b} \tag{3}$$

with

$$\mathbf{b} = (\mathbf{Z}'\Sigma^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\Sigma^{-1}\mathbf{y} \tag{4}$$

In Eqs. (3) and (4), **Z** and **y** are defined as in Eq. (2), y_0 represents the predicted response level, $\mathbf{z}_0$ represents the collocated sensor data (predictor variables), **b** represents the general-least-squares (Aitken) estimate of the $\boldsymbol{\beta}$ parameter vector, $\boldsymbol{\Sigma}$ represents the assumed (second-order stationary) residual spatial covariance structure, and **c** represents the assumed spatial covariance structure between the model residuals and the error associated with the new prediction site.

The LR model represents a special case of the SLR model; it arises naturally when the regression model residuals are assumed to be normally distributed and spatially independent (i.e., $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ and $\mathbf{c} = \mathbf{0}$). Thus, Eqs. (3) and (4) reduce to $y_0 = \mathbf{z}_0{}'\mathbf{b}$ and $\mathbf{b} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$, respectively. This model is actually just a “spatially referenced” multiple linear regression model where the model parameters correspond to various signal (and possibly additional trend surface and/or blocking) components.

2.2. Model estimation strategies

In most applied survey situations, the covariate signal data are typically not exhaustive (i.e., it is not collected absolutely everywhere). Strictly speaking, HSLR or CoK models should be used to analyze such data when inference is desired across all possible locations within the spatial domain. However, in practice the acquired grid data are often assumed to be sufficiently representative of the underlying spatial domain. In turn, this allows one to restrict his/her statistical inference to just the acquired grid data, and thereby simplify the modeling approach to a SLR (or KED) equation.

When the covariance structure is known up to a proportionality constant in the SLR model (i.e., $\mathbf{\Sigma} = \sigma^2\mathbf{V}$, where $\mathbf{V}$ is assumed known a priori), the $\boldsymbol{\beta}$ parameter vector in Eq. (2) can be estimated using generalized least squares (Graybill, 1976; Rao and Toutenburg, 1995). However, a specific $\mathbf{\Sigma}$ structure is rarely known a priori. In practice, the $\boldsymbol{\beta}$ parameter vector and $\mathbf{\Sigma}$ covariance structure must be jointly estimated from the sample data, typically using maximum likelihood (ML) or restricted maximum likelihood (REML) estimation techniques (Littell et al., 1996). In such situations it is generally necessary to collect a fairly large amount of sample data (>60 sites) in order to reasonably estimate the parameters associated with the covariance structure when even the simplest isotropic covariance structures are employed.

Unfortunately, this sampling load tends to be cost prohibitive in most commercial salinity assessment and/or precision farming applications. However, in many of these same survey situations the auxiliary sensor data are expected to be well correlated with the response variable of interest and the assumed residual error distribution is expected to exhibit only short-range spatial correlation. Under these conditions the simpler LR model can be used in place of the SLR model to generate a map with a high degree of prediction accuracy, provided that an appropriate sampling strategy is employed. This is especially advantageous in commercial applications, since LR models can be estimated using far less sample data (i.e., typically 10–15 sites). Furthermore, the LR parameter estimates and model predictions should be almost as efficient as the SLR parameter estimates and predictions provided that the model residuals approximately satisfy the spatial independence assumption (Kramer and Donninger, 1987).

In summary, there can be a substantial reduction in cost with only a minimal loss in precision when a LR model is used in place of a SLR or KED model, provided the simplier LR modeling assumptions are satisfied. In such a scenario, one would naturally like to employ a calibration sampling strategy that maximizes the possibility of ensuring spatial independence (with respect to the residual error distribution), while simultaneously choosing survey locations that in some way optimize the estimation of the $\boldsymbol{\beta}$ parameter vector. A prediction-based sampling strategy designed to achieve these two goals was introduced by Lesch et al. (1995b); an overview of this strategy is given in the next section.

3. Sampling strategies for the spatial linear model

3.1. Probability versus prediction based sampling plans

In general, probability-based sampling strategies are the most common type of sampling designs employed in spatial research problems. Probability-based sampling strategies include techniques like simple random sampling, stratified random sampling, cluster sampling, and multistage sampling for geostatistical estimation problems, and capture–recapture, line transect and adaptive sampling for detecting elusive populations, etc. (Thompson, 1992).

Probability-based sampling strategies have a well developed underlying theory and are clearly useful in many spatial applications (Thompson, 1992; Brus and de Gruijter, 1993). However, they are not designed specifically for estimating models, per se. Indeed, most probability sampling strategies explicitly avoid incorporating any parametric modeling assumptions, relying instead upon randomization principles (that are built into the design) for drawing statistical inference.

Prediction-based sampling strategies represent an alternative approach for developing sampling designs that are explicitly focused towards model estimation. The underlying theory behind this approach for finite population sampling and inference is discussed in detail in Valliant et al. (2000). More generally, response surface design theory and optimal experimental design theory represent two closely related statistical research areas that also study sampling designs specifically from the model estimation viewpoint (Myers and Montgomery, 2002; Atkinson and Donev, 1992). Techniques from the latter subject area have been applied to the optimal collection of spatial data by Müller (2001) and to the specification of optimal designs for variogram estimation by Müller and Zimmerman (1999). Conceptually, similar types of non-random sampling designs for variogram estimation have been introduced by Russo (1984), Bogaert and Russo (1999), and Warrick and Myers (1987).

The spatial response surface sampling approach discussed next represents a prediction (model) based sampling strategy. In this sampling approach, one assumes that some type of low order (linear or quadratic) regression model can be used to accurately approximate the soil property/sensor data relationship. The sample sites are then chosen to implicitly optimize the estimation of this model, subject to satisfying one or more explicit spatial optimization criteria.

3.2. Spatial response surface (SRS) sampling designs

To motivate the development of a spatial response surface (SRS) sampling design, assume that data from k sensors have been acquired at N representative survey sites across a field or survey region. Assume further that the soil property/sensor data relationship can be adequately described using a suitably specified linear regression model. In the SRS sampling approach, the goal is to select a small set of m sample sites ($m \ll N$) that serve to both (i) optimize the estimation of the regression parameters when using ordinary least squares estimation methods and (ii) eliminate or minimize the effects of the spatially dependent error structure on this estimation process.

The development of a SRS sampling design is a four step procedure. First, the acquired sensor data matrix (**Z**) is transformed into a centered, scaled, and de-correlated **X** matrix via a principal components analysis. Second, a traditional response surface sampling design is then selected that would in theory facilitate the optimal estimation of the regression model parameters associated with this **X** matrix. In the third step, an initial set of q "candidate" survey sites ($q > m$) are extracted from the **X** data matrix that most closely match the m design levels specified by the response surface design. Finally, a set of m sample sites are selected from the q candidate sites using an iterative algorithm that attempts to maximize some function of the minimum separation distance between adjacent site locations. This final step is performed in order to minimize (or eliminate) any short-range residual correlation effects and to facilitate the use of ordinary least squares estimation techniques. Some of the more pertinent details associated with each of these four steps are discussed below.

Step 1. The principal components transformation

Spatially acquired soil sensor data nearly always exhibit some degree of statistical correlation. Positively correlated data are especially common in both non-invasive and direct-contact soil conductivity surveys, due to the physical design of the electromagnetic induction (EM) and four-electrode systems. The primary reason for transforming the raw sensor data matrix (**Z**) into a standardized principal component data matrix (**X**) is to remove this correlation effect, as well as to standardize each score to have 0 mean and unit variance. (The sensor data should be standardized first, before the principal components analysis is performed.) This process is conceptually similar to the conversion of *natural* variables into *coded* variables in a traditional response surface model, and essentially facilitates the overlaying of a suitable response surface design onto the (transformed) sensor data.

As discussed in Lesch et al. (1995b), unusual sensor readings (i.e., outliers) are much easier to detect both visually and/or numerically after performing a principal components transformation. The principal components transformation aids in the preliminary screening of the sensor data for unreliable observations, in addition to facilitating the application of the response surface sampling design. However, unlike a typical principal components analysis, all of the principal component scores are normally retained during the site selection process.

Step 2. The choice of a specific response surface sampling design

The particular type of response surface design one should use in a specific survey application depends on both the assumed model and the number of sensors employed during the survey. When $k = 2$ or 3, second-order central composite sampling designs are often highly effective at minimizing the overall number of calibration sample sites, while still allowing for linear versus quadratic model discrimination. When data from four to six sensors are available, various hybrid and/or small composite designs can be quite useful. Fractional factorial designs can also be considered, if a first-order model can be safely assumed and the number of sensors becomes moderately large ($k \geq 6$). Detailed descriptions of these and other types of plausible response surface designs are given in Myers and Montgomery (2002).

Fig. 1 shows a second-order, rotatable central composite response surface design (CCRSD) overlaid onto a set of >7000 transformed and de-correlated EM38 (vertical and

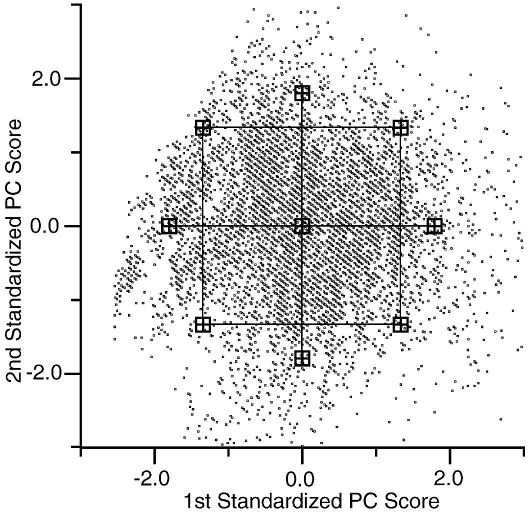

Fig. 1. A second-order CCRSD overlaid onto transformed and decorrelated EM38 sensor data, for purposes of identifying initial, candidate survey sites.

horizontal) sensor readings. The raw EM38 data were collected from a 64-ha alfalfa field located in Imperial, California, and are used here for illustrative purposes only.[1,2] In this example, approximately 80% of the bivariate principal component scores lie within a radius of 1.8 standard deviations from the (0, 0) mean, and the cube and axial points in the corresponding CCRSD have been adjusted to match this same radius. Hence, this design insures that approximately 80% of the survey data lie within the design region. Furthermore, although no specific (x_1, x_2) de-correlated survey readings exactly match any of the theoretical design levels, there are many sets of readings that fall quite close the each level (i.e., within a small tolerance distance).

Step 3. The initial selection of candidate survey sites

The selection of candidate sites represents a critical step in the optimization of the sampling design. In this step, a small number of initial sites are selected that closely match each specified SRS design level (i.e., the design points shown in Fig. 1). This selection process can be implemented in different ways, depending on the amount of survey data being analyzed.

In small surveys, two or three sites are normally selected for each specific design level based solely on their statistical distance from the design level coordinates. For example, suppose that N transformed principal component records are available from two sensors (k_1

[1] The limited EM38 signal resolution setting (±1 mS/m) has caused the diagonal banding effect that is apparent in the principal component data.

[2] Mention of trademark or proprietary products in this manuscript does not constitute a guarantee or warranty of the product by the U.S. Department of Agriculture and does not imply its approval to the exclusion of other products that may also be suitable.

and k_2), and two candidate sites are to be identified for each of m specific design levels. If the values of the jth design level are (τ_{1j}, τ_{2j}), the euclidean distance of the ith (k_{1i}, k_{2i}) data record to this design level is

$$\Delta_{ij} = \sqrt{(k_{1i} - \tau_{1j})^2 + (k_{2i} - \tau_{2j})^2} \tag{5}$$

Therefore, one would select the two sites that produce the minimum Δ_{ij} values for each of the m distinct design levels, yielding an initial set of $2m$ candidate sites.

In a large survey (i.e., $N > 1000$), the initial selection of candidate sites can often be substantially improved by using the above distance criterion (Δ) in conjunction with knowledge of the joint spatial position of the potential sites. For example, suppose q sites with Δ_{ij} values less than c_0 (for some small c_0 tolerance value) are considered potential candidate sites for the jth design level. Suppose also that the site having the smallest Δ_{ij} value is selected as the first candidate site, and that one additional candidate site is needed for this level. Recall that the goal of the SRS algorithm (to be described in detail in Step 4) is to maximize some function of the minimum separation distance between adjacent calibration sites. Thus, to help improve this maximization process, the second candidate site would be chosen (from the $q - 1$ remaining potential sites) such that the physical distance between the two candidate sites is made as large as possible.

The improvement gained from this type of combined minimum-statistical/maximum-physical distance selection criteria is shown in Fig. 2a versus Fig. 2b. Both sets of candidate sites were generated from Fig. 1 data, using the previously described CCRSD with two sites selected for each of the nine target design levels. The sites shown in Fig. 2a were chosen by selecting the two sites that minimized Eq. (5) at each design level. Fig. 2b shows how the candidate site coverage pattern changes when the second candidate site for each design level is chosen such that it is (a) within 0.15 standard deviations of the target design level and (b) exhibits maximum physical separation from the first candidate site (for that same design level). Although this technique only maximizes the minimum separation distances between candidate sites for distinct design levels individually (rather than all m design levels simultaneously), the improvement in the coverage pattern is still pronounced. In general, this approach tends to significantly improve the minimum separation distances between the joint set of initial candidate sites, particularly in large surveys.

Step 4. The final selection of an optimal set of sample sites

Once a suitable set of candidate sites have been identified, an iterative optimizing algorithm must be used to identify the “best” m sites, subject to some specific optimization criteria. Either parametric or non-parametric optimization criteria can be employed, depending on ones prior knowledge of the expected residual spatial correlation structure.

In the parametric approach, a hypothetical spatial covariance structure is assumed and a function of this covariance structure is directly minimized. For example, suppose that the residual error distribution is thought to be well approximated by an isotropic spherical model with range $\leq\alpha$, and define the optimization function to be the average correlation between the m sample sites. Let ε (the expected model errors) be distributed as $\varepsilon \sim N(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = \sigma^2\mathbf{V}$

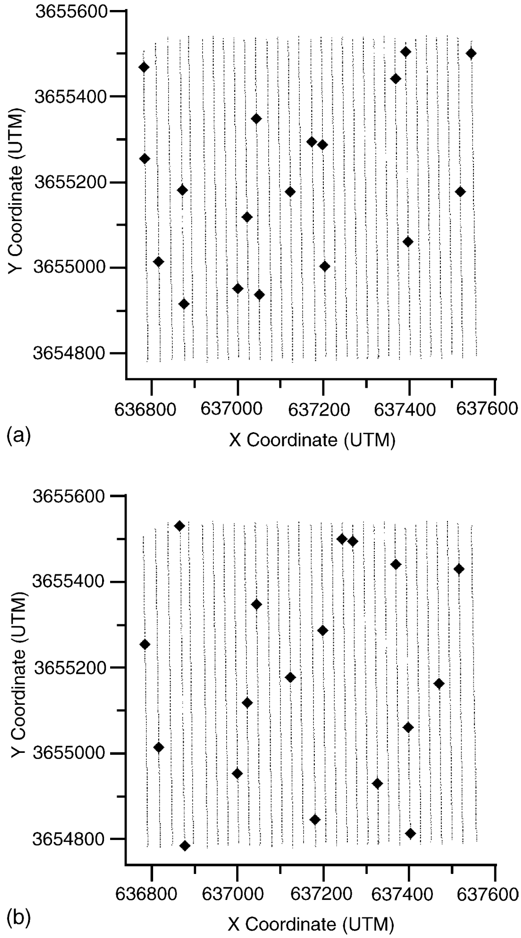

Fig. 2. Two sets of initial candidate sites from the same field chosen (a) without regard to the physical separation distances and (b) after maximizing the physical separation distance between sites within specific design levels.

represents the assumed covariance structure. The corresponding optimization criterion (φ) becomes

$$\varphi = \min\left(\frac{\mathrm{Var}(\mathbf{l}'\varepsilon)}{\sigma^2}\right) = \min(\mathbf{l}'\mathbf{V}\mathbf{l}) \tag{6}$$

where $\mathbf{l}$ represents a ($m \times l$) vector with each element equal to $l/m^{1/2}$. In the above example, if all of the m minimum separation distances exceed α, then $\mathbf{V} = \mathbf{I}$ and hence $\varphi = 1$ (which represents the absolute minimum obtainable value). Thus, as $\varphi \rightarrow 1$, the assumed sample site residual covariance structure approaches independence.

If $\mathbf{V}$ can be specified precisely, some type of alternative optimality criterion that directly minimized the expected variance–covariance matrix of the parameter estimates can instead be employed, e.g.,

$$\varphi = \max(|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|) \tag{7}$$

In Eq. (7), φ represents a generalized D-optimality criterion (Myers and Montgomery, 2002). However, such a criterion is only reasonable when the assumed covariance structure ($\mathbf{V}$) is known with a high degree of precision, since the use of Eq. (7) does not in general guarantee a final sampling plan where $\mathbf{V} \approx \mathbf{I}$.

In the non-parametric approach, a specific covariance structure is not assumed a priori. Rather, one only assumes that the correlation between sample sites will decrease as the minimum separation distance between sites increases. Thus, the algorithm attempts to maximize some suitable function of these separation distances. For example, let $\mathbf{d}$ and $\log(\mathbf{d})$ represent $(m \times 1)$ vectors of minimum separation distances and log distances for the current sample configuration, and $\mathbf{k}$ represent a $(m \times 1)$ vector of $\mathbf{l}$'s. Three suitable optimization criteria are

$$\varphi = \max\left(\frac{\mathbf{k}'\mathbf{d}}{m}\right) = \min\left(\frac{-(\mathbf{k}'\mathbf{d})}{m}\right) \tag{8}$$

$$\varphi = \max\left(\frac{\mathbf{k}' \cdot \log(\mathbf{d})}{m}\right) = \min\left(\frac{-(\mathbf{k}' \cdot \log(\mathbf{d}))}{m}\right) \tag{9}$$

$$\varphi = \max(\min(d_j)) = \min(\max(-d_j)) \tag{10}$$

Eq. (8) maximizes the average separation distance between sites, Eq. (9) maximizes the geometric mean separation distance, and Eq. (10) maximizes the absolute minimum observed separation distance value in the distance vector. Although all three criteria are reasonable, Eq. (9) tends to produce a good compromise between the average (Eq. (8)) and min–max (Eq. (10)) criterion functions for most survey applications.

Once an appropriate optimization criterion has been selected, a suitable search algorithm can then be employed to select the best set of final m sites from the larger set of c candidate sites (identified in Step 3). The simplest type of algorithm is a sequential search algorithm, which works as follows:

1. compute φ_0 for a starting set of m sites;
2. sequentially replace one of the current sites with an appropriately selected "swap-site" (chosen from the $c-m$ remaining sites), and recompute φ after each swap;
3. identify the site associated with the best (minimum) φ value computed during the $c-m$ swaps;
4. if $\varphi < \varphi_0$, exchange the two sites and return to Step 1 above, else declare convergence and exit.

This algorithm has the advantage of being very fast and simple to implement, although it does not necessarily guarantee convergence to a global optimum (with respect to the restricted set of initial candidate sites).

More complicated search routines can be employed, such as various types of simulated annealing algorithms (Krzanowski and Marriot, 1994). Also, it can sometimes be advantageous to perform a constrained optimization where a more traditional optimal experimental design objective is used in conjunction with one of the non-parametric distance criteria described above (Cook and Wong, 1994). For example, suppose that some of the initial candidate sites deviate significantly from their corresponding target design levels (this problem can occur in small samples, where there are only a limited number of survey sites to choose from). To protect against excessive degradation in the final design, a second criteria such as D-optimality can be simultaneously computed and tracked. If a particular site exchange results in a significant reduction in D-optimality, then it can be rejected, etc.

Regardless of the particular details employed in the search routine, the final goal of the SRS design is to select a set of *n* sample locations which are (i) nearly optimal with respect to matching the selected response surface design levels and (ii) physically separated far enough apart to ensure that the independent error assumption is adequately met. When a non-parametric optimization criteria is employed in Step 4, the resulting sampling locations tend to be spread nearly uniformly across the survey region. Under these conditions, the selection algorithm operates like a constrained space filling design (Müller, 2001, chapter 4).

3.3. Residual diagnostics for linear models

If a spatially referenced LR model is to be successfully used in place of a SLR model, then more restrictive modeling assumptions need to be met. A critical assumption in the LR model is residual independence. In a spatial context, this assumption implies that the residual errors exhibit a lack of spatial correlation, which is equivalent to independence under the additional assumption of normality. Brandsma and Ketellapper (1979) introduced a test statistic for detecting spatially correlated residuals, commonly referred to as the Moran residual test statistic (Haining, 1990; Tiefelsdorf, 2000; Upton and Fingleton, 1985). This test can be used to assess the adequacy of the residual independence assumption and is discussed in more detail in Appendix A.

Additionally, most well known residual analysis techniques used in an ordinary regression analysis are just as useful when applied to a spatially referenced LR model. These include techniques for assessing the assumption of residual normality (quantile–quantile plots), detecting outliers and/or high leverage points (plots of internally or externally studentized residuals), and detecting model specification bias (residual versus prediction plots, partial regression leverage plots, added variable plots, etc.). Methods or statistics for assessing the predictive capability of an ordinary regression model (such as the PRESS statistic) are also directly applicable to the spatially referenced model. Cook and Weisberg (1999) offer a good review of applied regression model diagnostics and assessment techniques.

4. A field salinity survey example

4.1. Field S2C: data modeling and analysis

The example survey data considered here are from a 1989 salinity survey of a 13.7-ha cotton field located in Kings County, California. The field was furrow irrigated, and the cotton crop had recently emerged when the survey was performed. The field consisted of sand and sandy loam soil types (coarse-loamy, mixed calcareous, thermic Aeric Haplaquent). Prior yield loss patterns suggested that localized zones of excessive salinity might be present (personal communication, land owner).

Both EM and four-electrode signal data were collected at 198 sites using a Geonics EM38 meter and Martek SCT-15 meter attached to a fixed surface-array electrode system, respectively. Hand-held horizontal (EM_H) and vertical (EM_V) EM38 signal readings were acquired at each site, in addition to two surface array conductivity readings ($W_{0–1}$ and $W_{0–2}$) corresponding to 0–30 cm and 0–60 cm sampling depths. The survey sites were located approximated 25 m apart. The exact (relative) coordinate positions were determined using a Zeiss DME theodolite system. Soil salinity (EC_e) samples were also collected at each of the 198 survey sites from 0 to 30 cm and 30 to 60 cm sample depths and measured on a saturation extract basis (Rhoades, 1996).

To simplify the statistical analysis, the salinity data have been averaged into composite 0–60 cm samples at each site. The EM_V signal data have also been discarded, since the target soil property is the near-surface salinity level and the EM_H, $W_{0–1}$, and $W_{0–2}$ signal data more effectively measure the 0–60 cm depth. All of the statistical calculations were performed using SAS Version 8 software (SAS Institute, 1999a,b).

Table 1 shows some basic summary statistics and quantile estimates for the complete set of sample salinity and instrument survey data. Equivalent statistics for the natural log transformed data are also given in Table 1, in addition to the ln(EC_e)/ln(instrument) correlation levels. The EC_e sample distribution is markedly right-skewed; most of the measurements (75%) fall below 3.4 dS/m, but the highest salinity levels exceed 16 dS/m. The sensor data distributions exhibit similar degrees of skewness and appear to be highly correlated with the EC data on the log scale. The apparent spatial ln(EC_e) pattern is shown in Fig. 3 and confirms that localized zones of salinization are present in the field.

Based on the exploratory data analysis, a linear signal+ trend surface regression model was specified to describe the log salinity/log signal data relationship across the 198 sample sites. The three log transformed signal readings (i.e., ln(EM_H), ln($W_{0–1}$), ln($W_{0–2}$)) were

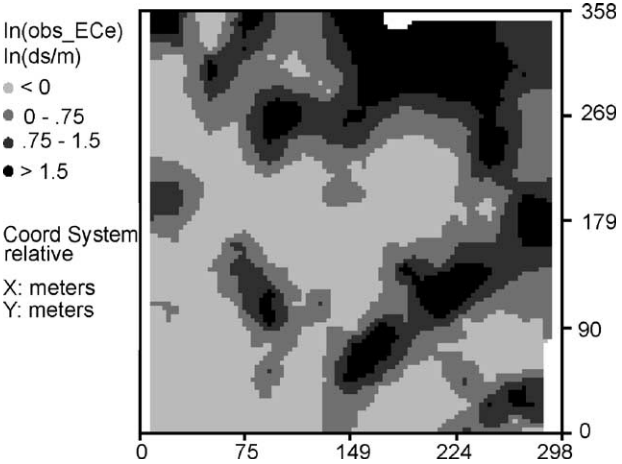

Fig. 3. Interpolated map of the observed log salinity pattern in Field S2C, based on 198 sample sites.

Table 1
Basic summary statistics for the Field S2C soil salinity and sensor data ($N = 198$)

Variable	EC_e (dS/m)	EM_H (mS/m)	$W_{0–1}$ (dS/m)	$W_{0–2}$ (dS/m)
Non-transformed sensor data				
Mean	2.86	53.60	2.04	5.53
Standard deviation	3.67	34.52	2.40	5.39
Standard error	0.26	2.45	0.17	0.38
Quantiles (%)				
100 (Maximum)	16.66	173.0	11.12	22.03
95	11.53	134.0	7.94	16.89
90	9.43	110.0	6.45	14.88
75	3.34	70.0	2.24	7.30
50 (Median)	1.03	42.0	0.93	2.92
25	0.64	28.0	0.62	1.79
10	0.46	23.0	0.48	1.34
5	0.36	18.5	0.39	1.00
0 (Minimum)	0.31	6.5	0.24	0.39
	$\ln(EC_e)$ (dS/m)	$\ln(EM_H)$ (mS/m)	$\ln(W_{0–1})$ (dS/m)	$\ln(W_{0–2})$ (dS/m)
Natural log-transformed data				
Mean	0.39	3.80	0.21	1.28
Standard deviation	1.10	0.60	0.94	0.92
Standard error	0.08	0.04	0.07	0.07
Calculated correlation matrix				
$\ln(EC_e)$	1.00	0.87	0.93	0.93
$\ln(EM_H)$		1.00	0.86	0.94
$\ln(W_{0–1})$			1.00	0.94
$\ln(W_{0–2})$				1.00

first standardized, and then a principal component analysis was performed to extract the three principal component scores (x_1, x_2, x_3). Additionally, the corresponding (u, v) survey site coordinates were redefined as $c_u = u/100$ and $c_v = v/100$. The initial regression model was then specified to be

$$\ln(\mathrm{EC_e}) = \beta_0 + \beta_1\, x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 c_u + \beta_5 c_v + \eta \tag{11}$$

where the model error (η) was assumed to be normally distributed with 0 mean and an unknown covariance structure ($\mathbf{\Sigma}$).

Eq. (11) was initially estimated under the assumption of residual independence ($\mathbf{\Sigma} = \sigma^2\mathbf{I}$), so that the regression residuals could be analyzed for evidence of spatial correlation and non-normality. This analysis confirmed that the normality assumption was reasonable, but the residual variogram plot revealed clear evidence of short-range isotropic spatial correlation (Fig. 4). Based on these findings, Eq. (11) was re-estimated via restricted maximum likelihood (SAS: Proc MIXED) using isotropic exponential, spherical, and gaussian covariance functions. Six distinct spatial covariance structures were actually used during the REML estimation process, since each covariance model was fit both with and without a nugget effect.

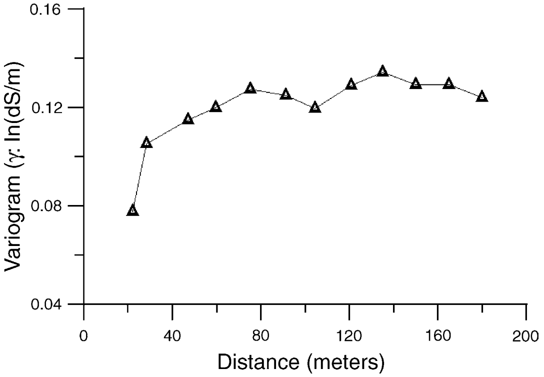

Fig. 4. Sample variogram plot of the OLS-LR model residuals.

Table 2 presents the −2 log-likelihood (−2LL) scores and relevant variance parameter estimates obtained from these analyses. As shown in Table 2, all six spatial covariance structures produced −2LL scores substantially lower than the residual independence model, confirming that residual errors are indeed spatially correlated. The exponential covariance function produced the smallest −2LL score among the no-nugget models, while the spherical function produced the smallest −2LL score among the models including a nugget term. All of the total sill estimates appear to be similar, as do the practical range estimates for four of the six spatial covariance structures. The appropriate nugget effect (if any) is more difficult to deduce from these results. The survey grid used in this study imposed a 25-m spacing between adjacent sample sites and hence the exact short-range spatial covariance structure cannot be effectively determined from these data.

Based solely on the −2LL scores, either the no-nugget exponential function or spherical function with a nugget effect can be judged to be the most appropriate covariance function for use with Eq. (11). Table 3 shows the corresponding regression model parameter estimates

Table 2
REML estimation results for Eq. (11): −2 log-likelihood scores and associated variance parameter estimates

Assumed $\boldsymbol{\Sigma}$ structure	Variance parameter estimates					
	−2LL[a]	Nugget	Partial sill	Range (m)	Effective range[b]	Total sill[c]
Spatially independent	179.0	0.126	n/a	n/a	n/a	n/a
Exponental, with nugget[d]	155.9	<.001	0.133	18.9	56.7	0.133
Exponential, no nugget	155.9	0	0.133	18.9	56.7	0.133
Spherical, with nugget	155.3	0.054	0.082	68.7	55.8	0.136
Spherical, no nugget	158.6	0	0.126	39.9	32.4	0.126
Gaussian, with nugget	155.7	0.067	0.066	31.7	54.9	0.133
Gaussian, no nugget	157.7	0	0.125	19.1	33.1	0.125

[a] −2 times the log-likelihood.
[b] The effective range is defined to be the separation distance at which the value of the estimated spatial covariance structure is 0.05 times the partial sill.
[c] Total sill = nugget + partial sill.
[d] Model converged, but Hessian was found to be not positive definite.

Table 3
REML regression model parameter estimates, standard errors, and *t*-test results for Eq. (11) under an assumed (a) spherical covariance structure (with a nugget effect) and (b) no-nugget exponential covariance structure

Variable	Estimate	Error	*t*-value	Prob (>*t*)
Under an assumed spherical covariance structure (with a nugget effect)				
Intercept	0.039	0.125	0.31	0.727
x_1	0.968	0.027	35.21	<.001
x_2	0.127	0.028	4.56	<.001
x_3	−0.044	0.024	−1.82	0.070
c_u	0.112	0.048	2.32	0.021
c_v	0.100	0.044	2.25	0.025
Under an assumed no-nugget exponential covariance structure				
Intercept	0.029	0.128	0.22	0.826
x_1	0.968	0.028	34.87	<.001
x_2	0.127	0.028	4.50	<.001
x_3	−0.039	0.024	−1.60	0.112
c_u	0.114	0.049	2.30	0.023
c_v	0.103	0.045	2.27	0.024

Estimates derived using all $N = 198$ observations.

and standard errors derived using each function. Both sets of estimates (parameters and standard errors) are nearly identical, regardless of the specific covariance function employed during the REML estimation process. Additionally, both sets of estimates show that the first principal component score represents by far the most important regression variable in the prediction equation. These results are consistent with the principal components analysis results, which found that the three corresponding eigenvalues explained 94.2%, 4.6%, and 1.2% of the observed variation in the log transformed and scaled signal data, respectively.

More importantly, the above results confirm that OLS estimation techniques should not be used to estimate Eq. (11), *assuming that all 198 sample sites can be used in the analysis*. However, a much smaller set of well chosen calibration sites could be combined with OLS estimation techniques to estimate Eq. (11) provided that a sufficient minimum separation distance is maintained between adjacent calibration site locations (about 57 m, based on the maximum practical range estimate shown in Table 2). Hence, a SRS sampling strategy would appear to represent a feasible approach for selecting a minimal set of efficient calibration sample site locations, assuming that cost constraints preclude the acquisition of a large number of samples.

To facilitate such an analysis, a rotatable second-order CCD having eight cube points, six axial points, and two center points was used to generate a 16-site sampling plan. The target design levels for this plan are shown in Table 4, along with some pertinent summary results generated by the SRS algorithm. Three candidate sites for each design level were initially identified using Eq. (5) (the euclidian distance function). A geometric mean separation distance criteria (Eq. (9)) was then employed to optimize the design. Fig. 5a and b shows the physical sample site locations of the initial and optimized designs, respectively.

The summary results given in Table 4 indicate that a fair amount of variation in the optimized design levels occurred during the optimization process (as compared to the target levels), but that the geometric mean separation distance was increased from an initial 48.1 m

Table 4
Summary results from the SRS sampling plan used in Field S2C

Second-order central composite response surface design levels ($n = 16$)

CCD			Final			Site #	MSD[a]
Target design levels			Optimized design levels				
x_1	x_2	x_3	x_1	x_2	x_3		
1.00	1.00	1.00	1.26	0.86	0.63	124	68.5
1.00	1.00	−1.00	0.37	1.28	−0.63	192	93.4
1.00	−1.00	1.00	0.64	−0.36	0.48	183	73.8
1.00	−1.00	−1.00	0.64	−0.78	−0.62	168	146.6
−1.00	1.00	1.00	−0.63	1.15	0.86	82	60.8
−1.00	1.00	−1.00	−0.86	0.90	−0.60	15	55.8
−1.00	−1.00	1.00	−1.26	−1.36	0.98	72	70.0
−1.00	−1.00	−1.00	−0.16	−1.92	−1.98	150	68.5
1.73	0.00	0.00	1.63	0.23	0.00	1	96.0
−1.73	0.00	0.00	−1.48	0.31	0.39	20	55.8
0.00	1.73	0.00	−0.59	1.75	−0.15	188	75.4
0.00	−1.73	0.00	−0.11	−1.30	0.19	68	69.6
0.00	0.00	1.73	−0.84	−0.10	1.45	25	68.2
0.00	0.00	−1.73	−0.33	−0.50	−1.34	28	68.2
0.00	0.00	0.00	0.09	0.31	−0.39	114	60.8
0.00	0.00	0.00	−0.21	0.02	−0.24	65	69.6

Geometric mean separation distances (m)	
In initial sampling design	48.1
In optimized (final) sampling design	72.8
Between site correlation (under assumed spatial covariance functions)	
REML spherical structure (Table 3(a))	
Maximum correlation estimate	0.053
Average correlation estimate	0.004
REML exponential structure (Table 3(b))	
Maximum correlation estimate	0.052
Average correlation estimate	0.003

[a] Minimum separation distance between sample site and nearest sample neighbor.

to 72.8 m. The smallest minimum separation distance (MSD) observed in the optimized design is 55.8 m, and most of the MSD values exceed 68 m. Additionally, the average expected residual correlation values under the previously discussed spherical + nugget and no-nugget exponential covariance structures are less than 0.004, and the maximum expected values are less than 0.053.

The 16 calibration sites identified in Table 4 (and Fig. 5b) were then used to re-estimate Eq. (11) using ordinary least squares; the relevant parameter estimates for this model are shown in Table 5. This model produced an R^2 value of 0.867 and a MSE estimate of 0.182. The t-tests associated with the parameter estimates show that the x_2 (second) and x_3 (third) principal component scores are non-significant, but that at least one of the trend surface components is marginally significant. A general set of residual diagnostic plots suggested no problems with the model specification, outside of the relative need for the x_2 and x_3 principal component variables. The model residuals satisfied the normality assumption, and a Moran

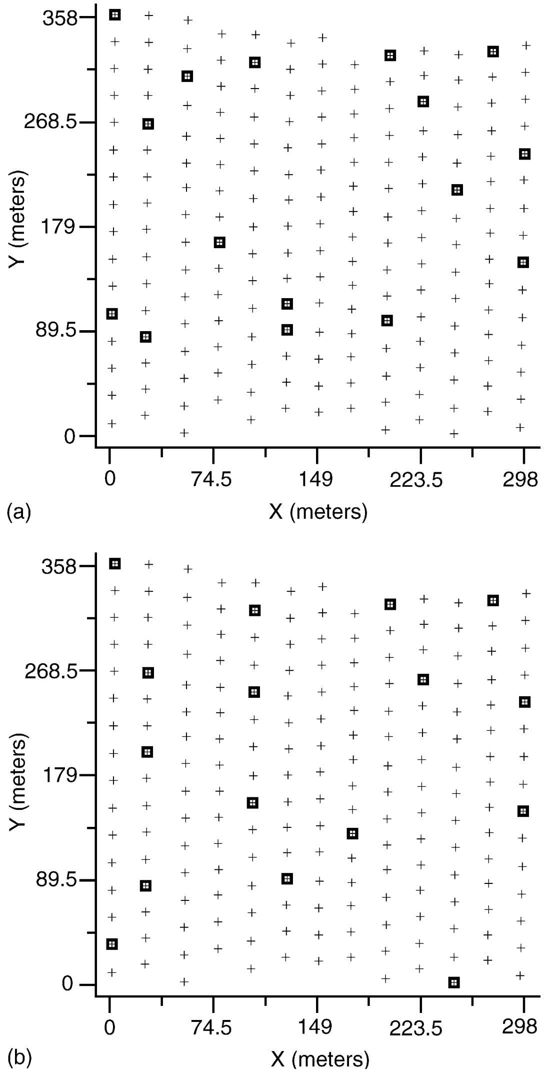

Fig. 5. Locations of the (a) initial and (b) optimized 16 calibration sampling locations for Field S2C, as determined by the spatial RSD algorithm.

test (for residual spatial correlation) produced a non-significant value of 0.73 ($p = 0.234$), suggesting that the assumption of approximate residual independence is reasonable.

Table 6 compares Table 5 OLS parameter estimates with the corresponding 16 site, generalized least squares (GLS) estimates derived using each of the previously discussed spatial covariance structures. The changes are trivial; there is virtually no difference between the three sets of calculated parameter estimates. Overall, there has been almost no loss in efficiency from using OLS estimation techniques in place of GLS estimation techniques in this example, because the residual independence assumption has been adequately met.

Table 5
OLS regression model parameter estimates, standard errors, and t-test results for Eq. (11)

Variable	Estimate	Error	t-value	Prob (>t)
Intercept	−0.178	0.336	−0.53	0.607
x_1	0.826	0.159	5.21	<.001
x_2	0.002	0.113	0.02	0.985
x_3	0.111	0.134	0.83	0.428
c_u	0.223	0.112	2.00	0.074
c_v	0.143	0.117	1.23	0.249

$n = 16$; Residual errors assumed to be spatially independent.

Table 6
Comparison of OLS and GLS regression model parameter estimates for Eq. (11)

Variable	OLS: $\mathbf{\Sigma} = \sigma^2\mathbf{I}$[a]	GLS: $\mathbf{\Sigma}$ = spherical model + nugget	GLS: $\mathbf{\Sigma}$ = exponential model, no nugget
	Estimate	Estimate	Estimate
Intercept	−0.178	−0.179	−0.178
x_1	0.826	0.827	0.827
x_2	0.002	0.002	0.002
x_3	0.111	0.111	0.111
c_u	0.223	0.224	0.224
c_v	0.143	0.143	0.143

$n = 16$.

[a] From Table 5.

The 182 observed versus predicted ln(EC$_e$) observations generated by the LR model are shown in Fig. 6. The calculated correlation between these observed and predicted ln(EC$_e$) values is 0.918. As shown in Table 7, the average calculated difference (error) between these data was −0.054, with an observed standard deviation and standard error of 0.44 and 0.033, respectively. The final predicted ln(EC$_e$) field map is shown in Fig. 7; this should be compared to the observed ln(EC$_e$) map shown previously in Fig. 3. In most respects, the predicted pattern agrees quite well with the observed pattern. The high saline zone in the

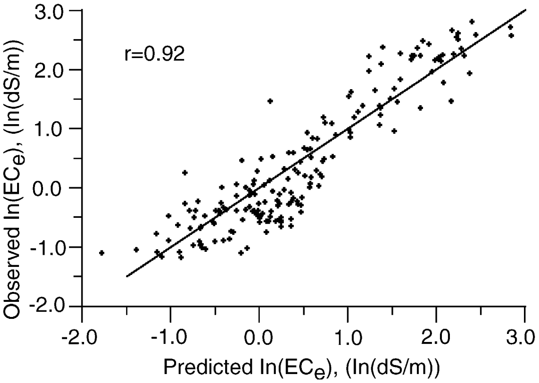

Fig. 6. Observed versus regression model predicted log salinity for the 182 validation sites, using predictions generated by the LR model estimated from the 16 calibration sample sites.

Table 7
Summary statistics for observed and predicted ln(EC$_e$) sample data (ln[dS/m]) for $N - n = 182$ prediction sites

	Observed ln(EC$_e$)	Predicted ln(EC$_e$)	Observed errors
Mean	0.395	0.449	−0.054
Standard deviation	1.110	0.963	0.444
Standard error	0.082	0.071	0.033

Prediction data generated using the OLS linear regression equation. Observed vs. predicted ln(ECe) correlation estimate: 0.918.

north-east section of the field has been correctly identified, in addition to the band of elevated salinity running diagonally across the south-east section. In general, this map is sufficiently accurate to allow for the effective implementation of a spatially referenced reclamation effort, and definitely accurate enough for overall salinity classification purposes.

Two points concerning this analysis warrent further discussion. First, in an actual survey process, the analyst will typically not have a priori knowledge of the correct underlying model. Instead, the analyst will have to use the limited set of calibration sample data to determine the appropriate model structure. Hence, employing a sampling design that aids in this effort clearly makes good modeling sense, and response surface designs tend to be some of the most efficient designs available for this purpose.

Second, the simple optimization algorithm described in Section 3 was successful at ensuring effective residual independence. Some of this success was gained at the expense of the final obtained design levels, which tended to exhibit a fair amount of deviation from their corresponding target levels in this optimized design. This latter effect is primarily due to the small total survey size ($N = 198$). In larger surveys this issue is rarely a problem.

4.2. Simulation results

As a final assessment of these survey data, a limited simulation analysis was carried out to compare some of the optimized SRS sample design features against traditional (i.e.,

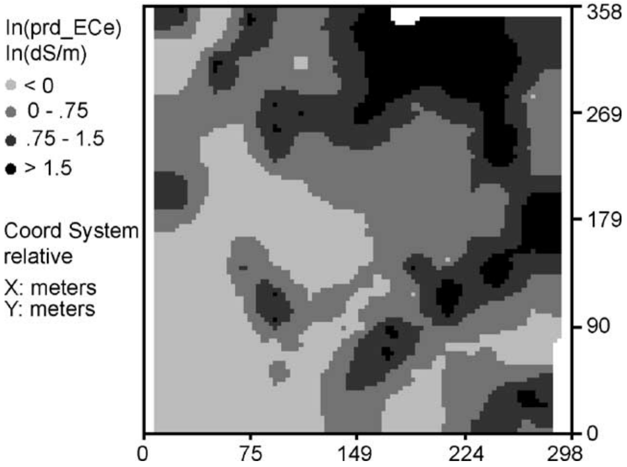

Fig. 7. Interpolated map of the LR model predicted log salinity pattern in Field S2C.

probabilistic) simple random sampling. In this simulation study, 5000 independent simple random sampling plans of size $n = 16$ were selected from the full ($N = 198$) survey data set. The following four statistical design criteria were then calculated for each plan:

1. the geometric mean separation distance (geoMSD), as defined in Eq. (9);
2. the average prediction variance (avePVar), defined as

$$\text{avePVar} = \frac{1}{182}\sum_{j=1}^{182}(1 + h_{jj}), \quad h_{jj} = \mathbf{x}'_j(\mathbf{X}'_j\mathbf{X})^{-1}\mathbf{x}_j \tag{12}$$

   where h_{jj} represents the corresponding hat leverage value for each new prediction, and $\mathbf{X}$ represents the matrix of response variables for Eq. (11) generated by the simple random sampling plan;
3. the maximum leverage value in the corresponding design matrix ($\max(h_{ii})$), defined as

$$\max(h_{ii}) = \max[\mathbf{x}'_i\mathbf{X}(\mathbf{x}'\mathbf{X})^{-1}\mathbf{x}_i, \quad i = 1, \ldots, 16] \tag{13}$$

   where the $\mathbf{X}$ matrix again represents the matrix of response variables generated by simple random sampling;
4. the statistical balance across the three principal components (Balance), defined as

$$\text{Balance} = \sqrt{(\mathbf{l}'\mathbf{x}_1)^2 + (\mathbf{l}'\mathbf{x}_2)^2 + (\mathbf{l}'\mathbf{x}_3)^2} \tag{14}$$

   where $\mathbf{x}_i$ represents the vector of ith principal component scores selected by the simple random sampling plan, and $\mathbf{l}$ is constant vector with each element equal to (1/16).

These four statistics measure important design criteria associated with the prediction (regression) model. The geoMSD quantifies the degree of adjacent-site separation distance achieved by the plan, while the avePVar quantifies the average increase in relative prediction variance across the remaining 182 prediction sites (under the assumption of residual independence). Likewise, the maximum leverage statistic represents the highest leverage site in the generated design, and the balance statistic measures the degree of balance in the design across the three principal component scores. In a robust sampling design, these latter three statistics should be as small as theoretically possible.

Table 8 shows the quantile statistics for these four statistical criteria calculated from 5000 simulated simple random sampling plans (for Field S2C), and compares these to the values obtained from the previously discussed SRS sampling plan. Recall that the SRS algorithm explicitly optimized the geoMSD criteria in the SRS sampling plan and achieved a geometric mean separation distance of 72.8 m. As shown in Table 8, the observed geoMSD statistics range from 27.6 to 67.1 m in the 5000 simulated plans, with 95% of the plans displaying values below 54.1 m and 50% of the plans displaying values below 43.8 m. None of the simple random sampling plans outperform the SRS plan with respect to this design criteria. Furthermore, the optimized SRS sampling design achieved about 29 m more adjacent-site separation than 50% of the simple random sampling plans.

The SRS sampling plan also performed quite well with respect to the latter three statistical criteria. This design exhibits an avePVar statistic that is better than about 89% of the simulated simple random sampling plans, and a Balance statistic that is better than

Table 8
Spatial response surface sampling plan vs. simple random sampling: simulation results ($N = 5000$ simulations)

	Design criterion (as defined in text)			
	geoMSD[a] (m)	avePVar[b]	Max(h_{ii})[c]	Balance[d]
Spatial response surface	72.8	1.47	0.672	0.134
Quatile estimates generated from 5000 simulated simple random sampling plans (%)				
100 (Maximum)	67.1	3.91	0.983	1.031
99	58.5	2.43	0.913	0.825
95	54.1	2.05	0.870	0.677
90	51.6	1.93	0.838	0.597
75	47.8	1.75	0.781	0.488
50 (Median)	43.7	1.62	0.712	0.373
25	39.9	1.53	0.647	0.268
10	36.9	1.47	0.602	0.191
5	35.1	1.43	0.574	0.149
1	31.9	1.39	0.538	0.088
0 (Minimum)	27.6	1.33	0.484	0.028
Number of simulated simple random sampling plans producing better design criterion values	0/5000 (0.0%)	553/5000 (11.1%)	1702/5000 (34.0%)	183/5000 (3.7%)

[a] Geometric mean separation distance: calculated using Eq. (9) in text.
[b] Average prediction variance: calculated using Eq. (12) in text.
[c] Maximum leverage value: calculated using Eq. (13) in text.
[d] Statistical balance across principal component scores: calculated using Eq. (14) in text.

about 96% of the simulated plans. It performed the least well with respect to the maximum leverage statistic, but still managed to do better than about 66% of the simulated plans. Perhaps most importantly, only 5 out of 5000 (0.1%) simple random sampling plans simultaneously produced smaller average prediction variance, maximum leverage, and balance values.

Although limited in scope, these simulation results are nonetheless informative. Specifically, these results show that by combining a response surface sampling plan with spatial optimization techniques, one can create a sampling design that is not achievable under a simple probabilistic sampling scheme. Equivalently, if a probabilistic sampling scheme is to be employed to make model-based inference, then a substantial number of randomization restrictions will need to be introduced in order to have any chance of generating a robust design with respect to these model-based criteria.

5. Conclusion

This paper has presented a unified sampling and modeling strategy for predicting soil property information from spatially referenced sensor data. Particular emphasis has been focused towards the fitting of spatially referenced regression models using ordinary least squares estimation and an objective, model-based sampling strategy designed to facilitate the estimation of such models. As with most sampling strategies, this approach possesses distinct advantages and disadvantages. The main advantages of this approach are two-fold. First, a substantial reduction in the number of samples required for effectively estimating a prediction equation can be achieved, in comparison to more traditional probability-based sampling designs. This reduction is achieved by adopting a prediction-based sampling approach, and employing a suitable response surface design to constrain the sampling algorithm. Through the selection of an appropriate (typically second-order) response surface design, uncertainty in the correct model specification can generally be minimized. Likewise, the likelihood of achieving approximate residual independence is greatly enhanced by maximizing the minimum separation distance between sample sites. A more complicated spatial linear model can normally be reduced to just a spatially referenced ordinary regression model by adopting this type of sampling approach, often with only a trivial loss in efficiency.

Second, this approach lends itself naturally to the analysis of remotely-sensed data. Ground, airborne, and/or satellite based remotely-sensed data are often collected specifically because one expects these data to correlate strongly with some parameter of interest (e.g., crop stress, soil type, soil salinity, etc.), but the exact parameter estimates (associated with the prediction equation) may still need to be determined via some type of site-specific sampling design. This approach explicitly optimizes this site selection process and does so in a highly cost-effective manner.

Disadvantages to this approach include the reliance on an a priori assumed model response structure, and the lack of any effective randomization in the site selection process. If the postulated model is grossly inadequate and/or the residual error structure exhibits a pronounced (long-range) spatial structure, then this sampling strategy will generally fail. These problems tend to occur most frequently when the sensor readings exhibit poor correlation with the target soil property and/or the target soil property is strongly influenced by secondary features that are not adequately measured by the sensor data. Of course, if the analyst suspects beforehand that the sensors will correlate poorly with the target soil property, then there is little point in developing any type of (probability or prediction based) sampling plan that conditions on the sensor response levels. However, if there is significant concern about the latter issue, then a more traditional probability based sampling strategy should generally be used in place of and/or in addition to the SRS sampling approach.

Appendix A. Abbreviations

EC	soil electrical conductivity
EC_a	apparent soil electrical conductivity
EC_e	electrical conductivity of the saturation extract
EM	electromagnetic induction
LR	(ordinary) linear regression
SLR	spatial linear regression
HSLR	hierarchical spatial linear regression

Appendix A (*Continued*)

UK	universal kriging
KED	kriging with external drift
CoK	cokriging
RSD	response surface design
SRS	spatial response surface
CCRSD	central composite response surface design
ML	maximum likelihood
REML	restricted maximum likelihood
−2LL	−2 log likelihood
$\mathbf{x}$	a vector
$\mathbf{X}$	a matrix
$\boldsymbol{\Sigma}$	a covariance matrix
$\boldsymbol{\beta}$	a vector of regression model parameters

Appendix B. The Moran test statistic

Brandsma and Ketellapper (1979) introduced a test statistic for detecting spatially correlated residuals, commonly referred to as the Moran residual test statistic (Haining, 1990; Tiefelsdorf, 2000; Upton and Fingleton, 1985). The Moran residual score (δ_{M}) is defined as

$$\delta_{\mathrm{M}} = \frac{\mathbf{e}'\mathbf{W}\mathbf{e}}{\mathbf{e}'\mathbf{e}} \tag{A.1}$$

where $\mathbf{e}$ represents the vector of observed model residuals and $\mathbf{W}$ represents a suitably specified proximity matrix. While the specification of $\mathbf{W}$ can be application-specific, in most precision agriculture applications it is generally reasonable to specify $\mathbf{W}$ as a scaled inverse distance squared matrix. Under such a specification, where d_{ij} represents the computed distance between the ith and jth sample locations, the $\{w_{ij}\}$ elements would be defined as:

$$w_{ii} = 0 \quad \text{and} \quad w_{ij} = \left(\frac{{d_{ij}}^{-2}}{\sum_{i=1}^{n} {d_{ij}}^{-2}} \right) \tag{A.2}$$

Brandsma and Ketellapper (1979) showed that the first two moments of δ_{M} are

$$E(\delta_{\mathrm{M}}) = \frac{\mathrm{tr}(\mathbf{MW})}{n-k} \tag{A.3}$$

$$\mathrm{Var}(\delta_\mathrm{M}) = \frac{\mathrm{tr}(\mathbf{MWMW}') + \mathrm{tr}(\mathbf{MWMW}) + \{\mathrm{tr}(\mathbf{MW})\}^2}{(n-k)(n-k+2)} - \{E(\delta_\mathrm{M})\}^2 \tag{A.4}$$

where $E(\,)$, Var( ), and tr( ) represent expectation, variance, and trace functions; n and k represent the number of sample sites (sample size) and regression model parameters (including the intercept); and $\mathbf{M}$ is defined to be $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The actual Moran test statistic is computed as

$$S_\mathrm{M} = \frac{\delta_\mathrm{M} - E(\delta_\mathrm{M})}{\sqrt{\mathrm{Var}(\delta_\mathrm{M})}} \tag{A.5}$$

where S_M is compared to the upper (one-sided) cumulative standard Normal probability density function.

When the Moran residual test statistic is found to be statistically significant, then the regression model residuals are said to exhibit significant spatial correlation. In this situation, the LR parameter estimates can be highly inefficient, the mean square error estimate and test statistics can be significantly biased, and the model predictions may be unreliable.

Appendix C. Assessing a strict residual independence assumption using the Moran residual test statistic

The previously described Moran residual test statistic represents a reasonable test for detecting spatially correlated residuals with respect to a specific sampling pattern. However, in certain situations the analyst may also wish to test for strict (i.e., absolute) residual independence. Strict residual independence in this context implies that the underlying error distribution is comprised entirely of pure white noise (or a pure nugget-effect, in geostatistical terminology), that these errors are spatially uncorrelated at all positive lag distances, and thus (given the additional assumption of residual normality) the regression model errors will be spatially independent regardless of the sampling pattern. In practice, if prior experimentation suggests that a particular sensor can accurately isolate and measure a specific soil property, then such an assumption might be tenable.

A test for strict residual independence can be performed provided duplication sample data is available. For example, consider the balanced case where two samples are acquired from each of n calibration sites (resulting in a total sample size of $2n$). Define the $\mathbf{W}$ matrix to be a binary (0/1) matrix were $w_{ij} = 1$ if samples i and j represent duplication sample data from the same calibration site, and $w_{ij} = 0$ otherwise. As discussed in Lesch et al. (1995b), the resulting Moran residual score (δ_M) then becomes a simple function of the traditional "lack-of-fit" test statistic (Myers and Montgomery, 2002). More specifically, under the assumption of strict independence and a correctly specified LR model,

$$g(\delta_\mathrm{M}) = \frac{(1+\delta_\mathrm{M})/n}{(1-\delta_\mathrm{M})/(n-k)} \; F_{n-k,n} \tag{A.6}$$

where F represents the F-distribution with $(n - k)$ and n degrees of freedom, respectively. Thus, if $g(\delta_\mathrm{M})$ is found to be statistically significant, the assumption of strict residual independence should be rejected.

References

Atkinson, A.C., Donev, A.N., 1992. Optimum Experimental Designs. Oxford University Press, Oxford, UK.

Bogaert, P., Russo, D., 1999. Optimal spatial sampling design for the estimation of the variogram based on a least squares approach. Water Resour. Res. 35, 1275–1289.

Brandsma, A.S., Ketellapper, R.H., 1979. Further evidence on alternative procedures for testing of spatial autocorrelation amonst regression disturbances. In: Bartels, C.P.A., Ketellapper, R.H. (Eds.), Exploratory and Explanatory Statistical Analysis of Spatial Data. Martinus Nijhoff, Boston, MA, USA, pp. 113–136.

Brus, D.J., de Gruijter, J.J., 1993. Design-based versus model-based estimates of spatial means: theory and application in environmental soil science. Environmetrics 4, 123–152.

Cook, R.D., Weisberg, S., 1999. Applied Regression including Computing and Graphics. John Wiley, New York, NY, USA.

Cook, R.D., Wong, K.W., 1994. On the equivalence of constrained and compound optimal designs. J. Am. Statistical Assoc. 89, 687–692.

Corwin, D.L., Lesch, S.M., 2003. Application of soil electrical conductivity to precision agriculture: theory, principles, and guidelines. Agron. J. 95, 455–471.

Cressie, N.A.C., 1991. Statistics for Spatial Data. John Wiley, New York, NY, USA.

Goldberger, A.S., 1962. Best linear unbiased prediction in the generalized linear model. J. Am. Statistical Assoc. 57, 369–375.

Graybill, F.A., 1976. Theory and Application of the Linear Model. Wadsworth Publishing Co., Inc., Belmont, CA, USA.

Haining, R., 1990. Spatial Data Analysis in the Social and Environmental Sciences. Cambridge University Press, Cambridge, UK.

Hendrickx, J.M.H., Das, B., Corwin, D.L., Wraith, J.M., Kachanoski, R.G., 2002. Indirect measurement of solute concentration. In: Dane, J.H., Topp, G.C. (Eds.), Methods of Soil Analysis, Part 4, Physical Methods. Soil Sci. Soc. Am. Book Series 5. Soil Science Society of America, Madison, WI, USA, pp. 1274–1306.

Kramer, W., Donninger, C., 1987. Spatial autocorrelation among errors and the relative efficiency of OLS in the linear regression model. J. Am. Statistical Assoc. 82, 577–579.

Krzanowski, W.J., Marriot, F.H.C., 1994. Multivariate Analysis. Part 1. Edward Arnold, London, UK.

Larson, W.E., Robert, P.C., 1991. Farming by soil. In: Lal, R., Pierce, F.J. (Eds.), Soil Management for Sustainability. Soil and Water Conservation Society, Ankeny, IA, USA, pp. 103–112.

Lesch, S.M., Corwin, D.L., 2003. Using the dual-pathway parallel conductance model to determine how different soil properties influence conductivity survey data. Agron. J. 95, 365–379.

Lesch, S.M., Rhoades, J.D., Corwin, D.L., 2000. ESAP-95 Version 2.10R: User Manual and Tutorial Guide. Research Rpt. 146. USDA-ARS, George E. Brown, Jr. Salinity Laboratory, Riverside, CA, USA.

Lesch, S.M., Strauss, D.J., Rhoades, J.D., 1995a. Spatial prediction of soil salinity using electromagnetic induction techniques: 1. Statistical prediction models: a comparison of multiple linear regression and cokriging. Water Resour. Res. 31, 373–386.

Lesch, S.M., Strauss, D.J., Rhoades, J.D., 1995b. Spatial prediction of soil salinity using electromagnetic induction techniques: 2. An efficient spatial sampling algorithm suitable for multiple linear regression model identification and estimation. Water Resour. Res. 31, 387–398.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., 1996. SAS System for Mixed Models. SAS Institute Inc., Cary, NC, USA.

Müller, W.G., 2001. Collecting Spatial Data: Optimum Design of Experiments for Random Fields, second ed. Physica-Verlag, Heidelberg, Germany.

Müller, W.G., Zimmerman, D.L., 1999. Optimal designs for variogram estimation. Environmetrics 10, 23–37.

Myers, R.H., Montgomery, D.C., 2002. Response Surface Methodology: Process and Product Optimization using Designed Experiments, second ed. John Wiley, New York, NY, USA.

Rao, C.R., Toutenburg, H., 1995. Linear Models: Least Squares and Alternatives. Springer-Verlag, New York, NY, USA.

Rhoades, J.D., 1996. Salinity: electrical conductivity and total dissolved solids. In: Sparks, D.L., (Ed.), Methods of Soil Analysis, Part 3, Chemical Methods. Soil Sci. Soc. Am. Book Series 5. Soil Science Society of America, Madison, WI, USA, pp. 417–436.

Rhoades, J.D., Chanduvi, F., Lesch, S.M., 1999. Soil Salinity Assessment: Methods and Interpretation of Electrical Conductivity Measurements. FAO Irrigation and Drainage Paper #57. Food and Agriculture Organization of the United Nations, Rome, Italy.

Rivoirard, J., 2002. On the structural link between variables in kriging with external drift. Mathematical Geol. 34, 797–808.

Royle, J.A., Berliner, M., 1999. A hierarchical approach to multivariate spatial modeling and prediction. J. Agric. Biol. Environ. Statistics 4, 29–56.

Russo, D., 1984. Design of an optimal sampling network for estimating the variogram. Soil Sci. Soc. Am. J. 48, 708–716.

SAS Institute Inc., 1999a. SAS/IML User's Guide, Version 8. Cary, NC, USA.

SAS Institute Inc., 1999b. SAS/STAT User's Guide, Version 8. Cary, NC, USA.

Thompson, S.K., 1992. Sampling. John Wiley, New York, NY, USA.

Tiefelsdorf, M., 2000. Modeling Spatial Processes: the identification and analysis of spatial relationships in regression residuals by means of Moran's I. Springer-Verlag, New York, NY, USA.

Upton, G., Fingleton, B., 1985. Spatial Data Analysis by Example. John Wiley, New York, NY, USA.

Valliant, R., Dorfman, A.H., Royall, R.M., 2000. Finite Population Sampling and Inference: A Prediction Approach. John Wiley, New York, NY, USA.

Wackernagel, H., 1998. Multivariate Geostatistics, second ed. Springer-Verlag, Berlin, Germany.

Warrick, A.W., Myers, D.E., 1987. Optimization of sampling locations for variogram calculations. Water Resour. Res. 23, 496–500.

Webster, R., Oliver, M.A., 2001. Geostatistics for Environmental Scientists. Johy Wiley, New York, NY, USA.